View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

经济学家也要学点网络爬虫技术

——漫谈爬虫技术与经济数据收集

厦门大学经济学院实验教学中心　钟锃光*

1　经济学实证研究中的的网络数据以及网络数据的特点

随着科技的发展，人们正面临信息爆炸。2010 年，零售巨头沃尔玛每小时都要处理 100 多万笔交易，为数据库大概上传 2,500 兆数据，相当于美国国会图书馆存书数的 167 倍。[①]可想而之，这个世界上数据量多到难以想象，而且还在不断地快速增长。

与此同时，经济学家的研究越来越离不开数据的支持。以 2012 年第 1 期的《经济研究》为例，11 篇学术文章，除了一篇纯理论研究的文章外，其余 10　篇均引用了各种数据。打开中国经济学工作者常常访问的论坛，也会发现里面有大量的信息是关于数据的下载、交换和交易的。为此经济学家不得不投入大量资金来购买各种数据库。

如果数据已经被很好的整理，即使需要高价购置，对经济学家而言已属幸运，实际上很多研究所需的数据往往无处寻觅。好在随着互联网的发展，电子商务、电子政务的逐渐推广，部分数据在网站上直接公开了，只是并未以良好的格式加以组织、对研究者不够友好。笔者在十年前做一个厦门市场房地产交易价格的论文时就碰到无法获取真实交易价格的问题，转而向多家在线房地产代理商索取数据，结果当然可想而知。被拒绝后，笔者对在线房地产代理商的网站进行了研究，并决定采用爬虫技术收集数据并最后完成了任务。

2 如何有效率地抓取网上数据－网络爬虫技术？

2.1　爬虫技术简介

爬虫是一种专门的程序，用于在互联网上自动抓取内容。最常见的爬虫是来自搜索引擎公司。在互联网刚刚兴起的 1994 年，yahoo 采用了层次归类的方法来索引当时的互联

*作者系厦门大学经济学院实验教学中心副主任兼王亚南经济研究院技术中心主任；电子邮箱：zhong.cn@gmail.com。

[①]Data, data everywhere，The Economist，Feb 25th 2010。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

网站点，在站点数目较少的时候，用手工还能处理不多的数据，随着互联网的发展，就需要自动化的工具来收集数据、更新内容、根据网站内容的链接来发现新的页面和网站，这时爬虫就变得必不可少了。早期的爬虫主要是索引网站中的文本内容，随着技术的发展，爬虫的功能也越来越强，例如对图片与内容的关联，对各种数据格式（如 pdf、doc、xls）的解析等。

在经济学研究中，其实并不需要像搜索引擎公司那样开发一种功能特别强大的爬虫，需要的是使用爬虫技术，能方便、大批量的下载网站上的数据，并且能够把数据整理成实证研究所需要的格式。

2.2 爬虫技术原理

用户在访问网页的时候，可能是打开某个网站作为起步，然后通过在浏览器上显示的该网站的内容，再加以浏览、点击等等，从而在不同的站点间跳转并获取信息。个人和服务器之间的交互以浏览器作为中介，浏览器把用户的点击，输入转化成 REQEUST（请求）并传输给服务器，服务器收到请求后，根据请求的内容，按需生成浏览器可以识别的数据格式，作为 RESPONSE(响应)传输给浏览器，浏览器解析服务器传递的内容，并把它显示成图文并茂的页面，这就完成了一次交互。可以看到，一次交互由 以下几步组成组成：

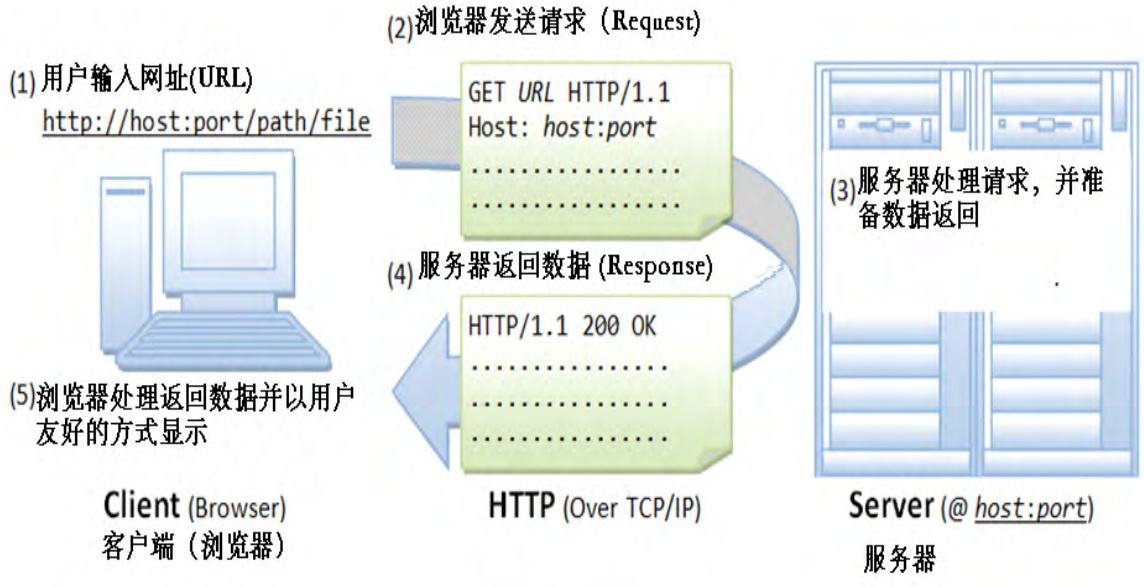

图 1：浏览器和服务器的一次交互

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一只标准的爬虫需要完成上述步骤中的 1,2,5 步，首先爬虫需要一个预先设置的起点，然后根据需要向服务器发送请求，这里的请求必需符合 HTTP 协议标准，在服务器看来，这个请求和正常浏览器发来的请求是一样的，所以照样生成相应的结果并返回给爬虫，这时爬虫收到的内容通常是 HTML 或其他浏览器可读的数据格式，但是爬虫不用显示这些内容，它需要解析这些内容，或保存，或丢弃，或者从里面发现其他的 Link (链接)来作为下一步的工作，这样一只爬虫就可以从一个起点，爬遍网上的每一个节点。

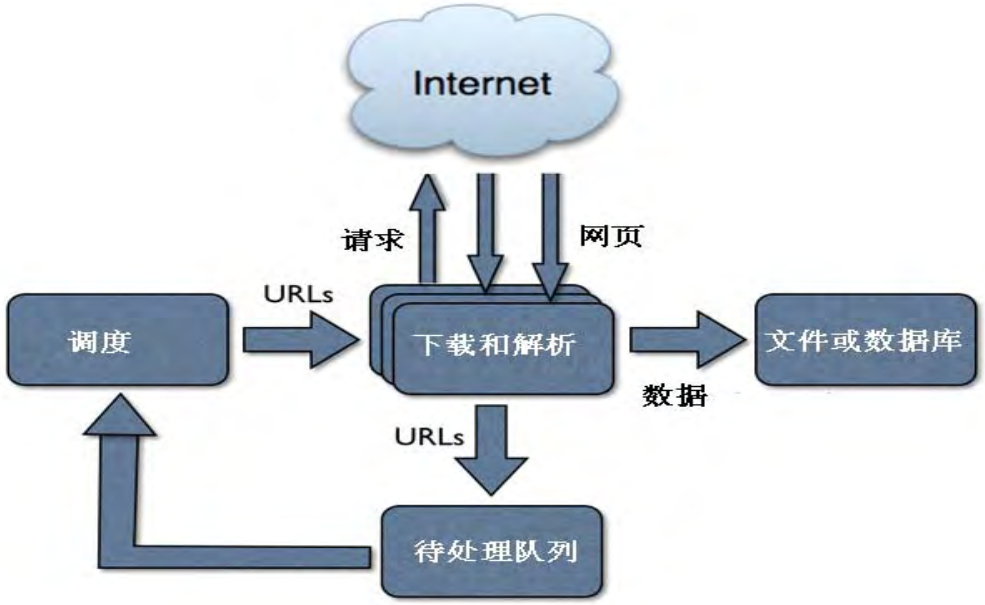

图 2： 爬虫工作原理

爬虫工作的关键一步是通过对服务器返回内容的解析得到想要的数据，只是这些数据原来是给浏览器准备的，对于普通人而已，那不是数据而是天书一样的乱码。服务器返回的数据有多种类型和格式，对于常见网页，服务器返回的是 HTML 格式，这是一种格式化的文本，和 Tex 有点类似。例如在浏览器地址栏输入 http://finance.yahoo.com/q?s=^gspc，浏览器将显示 S and P 500 的历史走势图，如果仔细观察可以在菜单中发现 Historical Prices，点击则进入历史数据显示页码，数据以网页表格形式呈现，更幸运的，最下面还有一个 Download to Spreadsheet 的链接，可以下载各类统计软件都可以直接使用 csv 格式数据。但是大多数时候没有这么幸运，如果不能直接下载 csv 格式，那么表格形式的网页也还算是不错的选择，只是要多费一些功夫，需要写一个 HTML Parser（HTML 解析器）

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

来分析网页，过滤掉网页代码中那些显示图片、动画、广告的代码，把需要的数据提取出来，保存为 csv 文件。此外还需要一个队列、调度来处理页面中的翻页行为。

2.3 学习爬虫技术的建议

爬虫说起来原理并不复杂，但是涉及到的面比较广。而且随着互联网技术的发展，一些新的技术不断出现，也需要爬虫能够跟得上技术发展的步法。

写一个爬虫，首先要理解 HTTP 协议和一些相关的知识，如 Javascript, HTML、XML、JSON 等，还可能根据具体的案例学习一些特殊知识。

其次还需要会一门顺手的语言工具，如果是简单的爬虫，SAS, R , MATLAB 等经济学家熟悉的语言都可以在一些扩展包的支持下完成简单工作，但是他们的扩展包相对于目前程序设计语言的中最流行几种语言[①]，那就少得可怜。根据在著名开源软件托管平台 github 和全球最热门程序员论坛 stack overflow 的调查统计，目前最流行的语言的前几名是 Java、JavaScript、PHP、C#、C++、Python、C，这些语言经过多年的发展，拥有大量的库函数，扩展包，第三方工具，只要选择学习其中的一门，有基本的编程基础，很快就可以在扩展库的帮助下写出第一个简单的爬虫。

最后就是需要不读的学习，勤思考，多动手，有问题及时通过搜索引擎查询相关知识库。

3 网络爬虫技术抓取网上数据的两个实例

3.1 抓取天气预报数据

项目组在一个研究中，需要用到天气的历史数据，经过 google 搜索，发现 http://www.wunderground.com 提供历史数据，在该网站首页的搜索框中，输入 xiamen，在返回的页面中发现了历史数据的链接："Weather History for 厦門市, Fujian"，点击后的页面显示了每小时的详细气候数据，这个时候可以发现到浏览器的地址栏显示：http://www.wunderground.com/history/airport/ZSAM/2014/3/23/DailyHistory.html

仔细观察这个地址，其中 ZSAM 代表了气象观察点代码，后面的 2014/3/23 则是时

[①]http://redmonk.com/sogrady/2014/01/22/language-rankings-1-14/; The RedMonk Programming Language Rankings: January 2014

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

间，把时间向前推，直到 1997 年 1 月 1 日都有精确到小时的记录。

http://www.wunderground.com/history/airport/ZSAM/1997/1/1/DailyHistory.html

这些气候信息在网页上以表格显示，本来还需要进一步的解析，幸运的在网页的底部还有一个连接“Comma Delimited File”，提供以逗号分隔的文本文件格式，也就是 CSV 格式，不在需要解析。为了下载带格式的数据，需要在地址后面多加个设置参数 format=1：

http://www.wunderground.com/history/airport/ZSAM/1997/1/1/DailyHistory.html?format=1

这时只需要构造一个简单的循环即可获得 1997–2014 年的厦门市历史气候信息。代码逻辑如下：

Loop 时间从 1997/1/1 开始到 2014/3/23

生成 URL http://..... /airport/ZSAM/时间/ DailyHistory.html?format=1

根据 url 下载对应的数据

合并保存到本地硬盘 csv 文件

End loop

如果要获得全中国的气候历史数据呢？那就需要更多的工作，从前面的工作可以看到 ZSAM 代表厦门，那什么代表福州？什么代表北京，是否这些城市都有历史数据呢？在该网站上可以发现一份完整的列表 http://www.wunderground.com/about/faq/international_cities.asp，其中有在中国的全部气象站列表，大部分的中国城市并没有拥有厦门这样形如 ZSAM 城市编号，用 5 位数字 WMO 编号的的才是普通规律，通过对 WMO 的查询，厦门气象一个更通用的 URL 应该是

http://www.wunderground.com/history/wmo/59102/2014/3/23/DailyHistory.html?format=1，其中 59102 表示城市的 WMO 编号，后面紧跟日期。改进后的获取全国历史气象数据的代码逻辑是：

分析气象站列表网页，获取全国气象站 WMO 编号列表

LOOP 中国气象站 WMO 列表

Loop 时间从 1997/1/1 开始到 2014/3/23

生成 URL http://..... /wmo/编码/时间/ DailyHistory.html?format=1

根据 url 下载对应的数据

合并保存到本地硬盘 csv 文件

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

End loop

END LOOP

这样一个中国历史气象数据的爬虫就完成了。这里只用到网页下载、字符串处理、文件读写等基本操作。读者可以尝试用 R 语言来实现它。

3.2 抓取房地产待售待租数据

一个研究项目需要利用北京现在二手房的挂牌价格，研究区域、朝向、学位、是否免税等因子与价格的相关性。首先选取了百度二手房网，http://esf.baidu.com ，该网站把所有房源分类成 大区、区域、小区、住房四个层次，每个住房标记了楼层、户型、价格、特点等，全部数据有 1 万个小区，133 万条记录，要下载分析这些数据绝非人力可为，更糟的是这个网站和上一例完全不一样，没有提供方便的下载方式，页面充满了复杂的广告、图片等、给数据下载工作带来很大的困难。

第一步还是要仔细观察网站了解数据的来源，在“按小区搜索”中，可以发现网站把北京分为 20 个大区，编号从 A1 到 A20，每个大区又分为若干区域，编号为 B0,B1...。网址 http://esf.baidu.com/bj/housecommunity/b0–a16/ 表示 大区 A16 顺义、b0 表示顺义的后沙峪，在返回结果有 35 个小区，分 2 页显示，页面参数分别是 n1,n2，最后每页 URL 是 http://esf.baidu.com/bj/housecommunity/b0–a16–n1/。通过对页面源代码的分析，可以看到每个小区有一个代码，例如双裕小区，代码 9926, 对应的房屋列表第二页 URL 则是 http://esf.baidu.com/bj/info/9926–10–n2 /。在这个页面可以直接获取房屋的相关信息。其逻辑是

LOOP 北京大区

LOOP 区域

LOOP 小区

LOOP 每页

分析 HTML DOM 并记录到数据库

END LOOP

END LOOP

END LOOP

END LOOP

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

可以看到，这个例子主要的困难在于对地址分析，代码和实际名称的对应，以及对每页数据的分析。这里需要用到 HTML DOM 解析，在 CPAN.ORG 上可以下载到 R 语言可用的 HTML–DOM 扩展包[①]。如果用 JAVA 或者 C#，则有更多的选择[②]。

4 网络爬虫的主要困难

从上一节的例子可以看到，使用爬虫下载数据，不仅需要技术，还需要细致和耐心。随着数据的重要性越来越被大家认识，对数据的公开的态度也分成两派，一派是提供各种数据下载、查询接口，甚至直接提供开发 API 方便大家使用，一派是设置各种障碍，防止数据被第三方使用。为此各种障碍技术被不断开发出来，例如根据爬虫的工作原理，一旦发现远端有规律地、大量地下载数据，就可以直接阻断连接。有的网站开发 flash 客户端，与服务器端用私有协议通讯；有的网站把数据处理成图片，让普通的文本解析器无法工作。在这种情况下，爬虫程序员不得不不断学习、试验各种新技术。

对于有意开发代码下载数据的经济学研究者，国内市场有两本书可以作为很好的教材，Michael Schrenk 著的《Webbots、Spiders 和 Screen Scrapers：技术解析与应用实践》，罗刚和王振东合著的《自己动手写网络爬虫》都比较详细的讲解了爬虫的开发、应用以及各种细节。至于所用到的 HTML、 Javascript、Json、web service 等相关知识，可以在开发过程中，根据数据的实际需要再了解相应的知识。

[①]http://search.cpan.org/~sprout/HTML-DOM-0.053/
[②]http://htmlparser.sourceforge.net/

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net